INL/CON-05-00065
PREPRINT

Planning for Quality Data

51st International Instrumentation Symposium

Robert P. Evans

May 2005

The INL is a
U.S. Department of Energy
National Laboratory
operated by
Battelle Energy Alliance

PLANNING FOR QUALITY DATA

Robert P. Evans
Idaho National Laboratory
Idaho Falls, Idaho 83415

KEYWORDS

Data, Quality, Uncertainty

ABSTRACT

The assurance of data quality can be a complex process requiring careful planning. The planning process described in this paper uses Data Quality Objectives as the foundation. The described process considers three steps: project requirement identification, definition of the information necessary to answer the questions, and data collection and management. Since sufficient levels of documentation are required at all levels, uncertainty analysis, traceability and custody, data maintenance, and data evaluation and review are also discussed.

INTRODUCTION

We live in an information age. We are constantly bombarded with information on various topics and from many different sources. It has been said that we suffer from information overload. [1] In the business world, information is becoming more critical as products become more complex, more expensive to develop, and competition becomes more intense to produce a better and cheaper product faster.

Generally, information is used to make decisions. In everyday life, information in the form of a buzzer helps us to decide to get up in the morning. Information from the gas gauge on the car helps us to decide that it is time to stop and fill the car with gas. In the same way, a business uses information obtained from sales reports and projections to make decisions on whether to open a new store or close one down. In an experiment, decisions are made based on information received from system measurements. The value or accuracy of the decision made is dependent on the accuracy of the information used.

At any level of management, decisions are made that affect the quality of the product and the company's ability to keep customers and attract new ones. Before these decisions can be made, information must be obtained, the quality of which will directly affect the quality of the decisions made. This points to the need for a systematic approach to the information-gathering portion of the decision process.

The need to provide information cheaper and faster often conflicts with the requirement that information collected be collected in a defensible manner. This conflict can overwhelm a project to the extent that no goals are met. Systematic planning can result in decisions being made based on data of known quality, saving both time and money. Accurate information is not normally obtained by chance. It takes careful planning and study to have the correct information for the decision to be made. It has been said if one fails to plan, they should plan to fail. Within the research environment, information, in the form of data, is being sought to make decisions or to evaluate the performance of a particular process. Lack of planning often leads to data whose quality is below the level needed, so that decisions made on the basis of the data may be incorrect. To avoid such difficulties, a data quality planning process must be implemented for any data collection effort. This paper presents such a process which can be used to obtain data of the desired level of quality.

DATA QUALITY PLANNING

Data quality is achieved and maintained by inclusion of several necessary elements in planning. Typically, Data Management Plans include:

- Data Quality Objectives
- Uncertainty Analysis
- Data Traceability/Custody
- Data Evaluation and Review
- Data Maintenance

Before launching into the data quality planning process, it would be best to define some terms as they are used in this paper.

Data: Information obtained and organized for the purpose of making decisions concerning the questions at hand. They are used to support a theory, prove a process, or verify a condition

Quality Data: Data that are of known and sufficient accuracy to answer the questions at hand.

Data can be obtained through many sources. These include literature searches, analysis, and testing. In each of these cases, a 'dependability' number can be assigned to the accuracy of the data. As an example, technical information received from the local daily newspaper is probably not as reliable as that obtained from a reviewed technical journal. In an analysis, the accuracy of the result can never exceed the accuracy of the input data. The accuracy of the calculation of the area of a circle will never be greater than the lesser of the accuracy of the value of 'π' used or the accuracy of the radius measurement. This concept of accuracy will be discussed further in the Uncertainty Analysis section of this paper. The accuracy of data obtained through testing is dependent on many factors including the accuracy of the measurement device, the principle of the test, and the test environment. Each of these must be addressed in order to understand the quality of the data obtained.

Although data quality is dependent on accuracy, it is not the same as accuracy. Using the above definition of quality data, the data received from a dime store thermometer, with an accuracy of 5 degrees, used to make a decision of whether it is warm enough to go swimming may be of a higher

quality level than the data received from an expensive temperature measurement system with an accuracy of 0.1 degree used to control a process requiring precision of 0.05 degrees. Data quality is based on the requirements of the project.

The need for quality data is best understood by considering the consequences of decisions based on poor quality data. Consider again the gas gauge on a car. An inaccurate gauge could leave a person stranded when the car ran out of gas even though the gauge indicated there was still gas in the tank. With inaccurate data, a business might build a new store in the wrong location and end up going out of business. A classic engineering example is the Tacoma Narrows Bridge connecting the Olympic Peninsula with the Washington State mainland. Although data had been taken to determine wind loads, the data was not of sufficient quality, therefore some bad decisions were made during the design and construction. The end result was the failure of the bridge.[2]

The achievement of quality data does not happen by accident. It is dependent on the identification of the needs and requirements of the project. In general, this requires a formal, standardized process to ensure that all pertinent questions are addressed. Although not a unique process, the data quality objective process does supply the needed methodology to accomplish these goals.

THE DATA QUALITY OBJECTIVE PROCESS

The data quality objective (DQO) process is a three-step process by which the objectives can be made clear, and uncertainties that can be tolerated in the context of the decision or objective are defined. The DQO process was originally defined by the Environmental Protection Agency [3] for use in remedial investigations; however, it can be adapted to any data gathering situation.

The first step of the DQO process requires input of the project requirements. In this step it is necessary to identify and clearly state the questions to be answered, the decisions to be made, or the objectives of the testing.

The second step is to define what information is required to answer the question, make decisions, or test objectives defined in the first step.

Step three is the data collection process. In this step, the information identified in the second step is collected and organized into a form such that the questions from the first step are clearly answered.

These three steps are discussed in greater detail.

Step 1 – Project Requirement Identification -

- Specify objectives/decisions
- Identify customer
- Identify available information
- Develop/describe conceptual model.

Requirements are a statement of the wants and needs requested by the customer. Sometimes these requirements are in the form of a formal, well-defined document, but too often it is up to the investigator to ferret out the requirements. Even in cases where there are written requirements, they are often incomplete or ambiguous. It is therefore important, in order to determine what the customer really wants and needs, to ask some questions. These might include:

- What is the problem?
- What does the customer want?
- What does the customer need?
- Why does the customer want it?
- What is the benefit to the customer?
- How will customers know when they have what they want?
- What questions are to be answered?
- What decisions will be made based on the information?
- What are the consequences of an incorrect decision?
- What current information is available on this question?

Requirements are not a solution to a problem but are usually a statement of the problem. It is not a specification but a question.

This first stage can be conducted in several ways, such as meetings, requests for information, or one-on-one interviews with the requester or those that will be the end user of the final product.

Step 2 – Define the information necessary to answer the questions

- Identify data types
- Identify data users (customers)
- Identify and specify data quality and quantity needs
- Evaluate available data
- Identify schedule and budget constraints
- Review data quality assigned
- Identify sampling design options
- Identify accuracy/uncertainty requirements.

Defining the information necessary to answer the questions requires an in-depth understanding of the question. This requires a second set of questions that are addressed to experts in the field of knowledge. These questions might include:

- What physical principles are involved?
- What experiments might provide the answer?
- How accurately do the answers need to be known?
- What quantities must be measured and how accurately?
- What variables must be controlled?

This step normally has at least two phases: 1) examination of the current base of knowledge to determine if others have addressed the same or similar problems; this is essentially a literature search, and 2) an examination of the physical principles involved and what physical parameters must be determined to answer the question.

The first phase is fairly straight forward. In the second phase, all of the parameters that supply input to the question are identified, and how they affect the answer is defined. At this point it is also necessary to determine how accurate the answer must be to answer the original question. From this analysis, those parameters which have the greatest effect can be identified and a determination made as to the required accuracy of each parameter in order to obtain the overall accuracy required. Determining the level of uncertainty allowable in the data may be an involved process. The total uncertainty of the system must be considered including those involved with the physical process, data collection uncertainties, sampling uncertainties, processing and conversion uncertainties, and any associated errors. This will be addressed in greater detail with the discussion of uncertainty analysis.

As a part of this second step, levels of data assurance should be assigned and a preliminary uncertainty analyses performed. These are determined on the basis of how the data will be used and the questions to be answered. Questions involving risk to life and property or those answering legal questions will require a higher level of data assurance than a question providing scoping information. Three levels are discussed along with data planning requirements:

Level 1: Legal or Technical – Complete data planning process, combined with documentation organized under a configuration management plan and with independent review. Level 1 data will be legally and technically defensible. These data might be used in courts of law or in making decisions concerning technical principles involving large commitments of time or money.

Level 2: Technical – Complete data planning process, combined with documentation organized under a quality program plan, with peer review. Level 2 data will be technically defensible. These data might be presented in technical publications.

Level 3: Scoping – A formal data planning process is not required, although an informal process is still recommended. This level allows for data for which quality requirements are not as rigorous. These data might be used to determine if a setup is going to provide "ball-park" information prior to proceeding with a higher level test, providing proof-of-principle, or in making design choices.

Although using a higher quality level in obtaining data is always safe, it does increase the costs and time involved. It is therefore important to determine how the data will be used prior to beginning the project.

Step 3 – Data Collection

- Perform uncertainty analysis / estimates
- Experimental design matrix rational
- Identify documentation required
- Identify data collection process

- Prepare DQO document.

In the third step of the DQO process, the data collection program is designed using as input the output of the first two steps. Designing the data collection program will include designing a measurement system and sampling strategy and deciding what documentation will be used to support the data. In step 2 the total uncertainty that could be tolerated in the program was decided and scoping uncertainty estimates determined. This is now used as input to determine the types of transducers, signal conditioning, data collection and data processing equipment to be used. Once each of these components have been identified, a formal, detailed uncertainty analysis can be performed on the entire system to determine if it still falls within the requirements envelope.

This step also includes the actual collection of the data. Care must be exercised during this phase to ensure that errors are not introduced and that the data are collected under the specified conditions. Care must also be taken to ensure that the data are not corrupted following the data collection process.

There are other details, beyond the scope of this analysis that must be taken into account in the data collection process. These include such things as instrument and system calibration, preparation of test plans and procedures, system interface requirements, system drawings and diagrams, wiring diagrams, installation procedures, and test setup and start-up procedures.

UNCERTAINTY ANALYSIS

Uncertainty is an expression of the difference between the measured value and the true value.[4] This difference is composed of two components: bias error (fixed or systematic) and random error (precision). These uncertainties may be expressed as an interval in which there is confidence the true value may be found or as a percentage of the measurement reading or the measurement range. The theory and detailed processes associated with the uncertainty analysis process is beyond the scope of this presentation.

An uncertainty analysis is performed in three stages: 1) Scoping, 2) Pre-measurement, and 3) Post-measurement.

Scoping – The scoping uncertainty analysis is performed during step 2 of the DQO process. The system, including the physical process, measurement system (transducers, signal conditioning, data collection), and data analysis, is analyzed in general terms in order to identify areas of large potential uncertainty and to determine if the "ball-park" uncertainty will allow the data to be used to answer the questions posed. By determining areas of greatest potential uncertainty, more effort can be placed in these areas to reduce the total uncertainty.

Pre-measurement – The pre-measurement uncertainty is performed prior to making measurements but after the final design of the measurement system. This is a detailed uncertainty analysis using the calculated uncertainties for each of the system components. This analysis is to determine if the final pre-measurement uncertainty is still within the envelope necessary to obtain useful information. At this point it is still not too late to rework the system without incurring large costs.

Post-measurement – The post-measurement uncertainty analysis is much like the pre-measurement analysis except now there is real information which can be analyzed as part of the uncertainty equation. The outcome of this analysis will be used to determine the uncertainty bands which are reported with the information.

Each of the stages of the uncertainty analysis are performed in the same manner, each step going into more detail and more rigor. Uncertainty analyses for a particular measurement system will be different than for another measurement system and hence each analysis must address specific possibilities for error in the system under analysis.

Although the specifics for each analysis may be different, the method will be much the same. It has been divided into steps below.

The first step of the process is to identify the measurement and the physical principle involved in the measurement. If more than one principle is involved, all must be identified. This principle will help to determine how the data will be used to determine the answer to the problem at hand. It is very seldom that the direct measurement itself is the answer to the question. Normally one or more measurements will be combined to determine the information required.

The second step is to determine how the measurements will be made. This is a point to look at tradeoffs in cost and accuracy. The analysis will give an indication how critical the accuracy of a particular measurement is in determining the accuracy of the system as a whole.

Step three is to determine the uncertainties of each of the components in the system. In the initial stages of the analysis process, this may be performed by estimation. In the latter stages, the information is obtained through consultation of vendor data or testing. Most of these will have several levels of uncertainty associated with them. For example, a transducer will have uncertainties associated with the calibration which in turn has uncertainties associated with traceability to known standards, calibration methodology, calibration laboratory conditions, etc. Each of these inputs must be considered, but to different degrees of rigor depending at which stage the analysis is performed.

In step four, the uncertainties are combined using approved and documented methodologies to obtain a total measurement system uncertainty.

POST TEST DATA HANDLING

Like the post test uncertainty analysis, the following items are not really part of the planning process but must be considered if quality data are to be obtained. These items include: 1) Data Traceability and Custody, 2) Data Evaluation and Review, and 3) Data Maintenance. Data Traceability and Custody and Data Maintenance are more important when dealing with legal or technical data requirements than when dealing with technical or scoping data because it may be necessary to verify that the data displayed or used is the same data produced from the testing. In all cases it is important to evaluate and review data before it is used in the decision-making process. Data Traceability and Custody and Data Maintenance will be presented here for completeness but they will not be discussed in detail. The planning process must still consider these items if the information required has legal constraints.

DATA TRACEABILITY/CUSTODY

Data traceability is used to describe the documentation supporting a logical path back to an accepted standard or practice. In calibration, for example, the accepted standard is the National Institute of Standards and Technology or a similar organization which holds and maintains reference standards for certain fundamental quantities.

Custody refers to the ability to account for the physical location and person in control of information or a sample to ensure that it was not left so that it could be changed or altered. For cases in which data may be used in a court case, lack of proof of custody may result in the data being legally inadmissible. In general, documentation should be available that shows that the data in question have been accumulated in compliance with standards or in standard practices or procedures that govern the data.

In order to provide for both traceability and custody, the planning process must include provisions for these items.

DATA MAINTENANCE

Data maintenance is designed to ensure that the original data that were collected have not been altered while processing, updating, or correcting data. Data maintenance procedures should be documented.

DATA EVALUATION AND REVIEW

Before data are reported or used to make a decision, they should be evaluated. Data validation is a systematic process for reviewing data against a set of criteria to provide assurance that the data are adequate for their intended use. The level of assurance chosen in the planning stage will affect the manner in which the evaluation is conducted. The following steps are part of evaluating the data:

1. Assemble the required documentation: The documentation trail for the data should have been part of the planning process.

2. Procedure compliance check: This is the process of verifying that all required procedures were followed and all planning requirements have been fulfilled.

3. Data review: A review of the data against a set of criteria. These criteria may be such things as known physical parameters at a given point in the test process. Three levels of data assurance correspond to three validation levels:

    - Level 1 – Data are validated and qualified by an independent review committee, which reviews the entire planning process, followed by a review of the data. They will also verify that the assigned uncertainties are correct.

- Level 2 – Data are validated and qualified by a project review committee. There is not the same level of rigor in the review of planning or the assigned uncertainties as was used in Level 1. The data are reviewed by methods appropriate to the data being reviewed.

- Level 3 – Data are subjected to peer review. This process may use any of the review processes above but lacks the independence and rigor of Level 1 or Level 2.

Three validation designations [5] that may be applied to any of the above levels of data validation are:

QUALIFIED - Data determined to represent the phenomena being measured and to meet the established objectives.

TREND - Data do not fully represent the phenomena being measured and/or do not satisfy objective requirements, but do contain useful information.

FAILED - Data contain no useful information relative to the measurement or to requirements.

Data that have not been evaluated are designated as "not reviewed."

CONCLUSIONS

Good information is necessary in order to make good decisions, without which an organization or business can find itself in deep trouble. Quality data can be achieved through careful planning and the use of Data Quality Objectives. This paper provides several general steps to assist in the planning and the evaluation process that can lead to data that are not only quality data but can be demonstrated to be quality data.

ACKNOWLEDGEMENTS

The information presented in this paper is the result of years of experience of measurement systems engineers at the Idaho National Laboratory gained through programs such as the Loss of Fluid Test program, the Semiscale Test Program, the Power Burst Test Program, and other small test programs. This paper is an attempt to document this information in a formal manner for others to use. The author acknowledges the contribution of many people that developed and tried these concepts over several years.

REFERENCES

1. Krill, Paul, *Overcoming Information Overload*, InfoWorld, January 7, 2000, http://www.infoworld.com/articles/ca/xml/00/01/10/000110caoverload.html

2. Petroski, Henry, *To Engineer is Human*, Vintage Books, New York, 1992, pp. 163-169

3. *Data Quality Objectives for Remedial Response Activities*, EPA/540/G-87/004, U.S. Environmental Protection Agency, Washington, DC, March 1987

4. Dieck, Ronald H., *Measurement Uncertainty, Methods and Applications*, Instrument Society of America, Research Triangle Park, NC, 1992, pp. 9-31

5. *Standard Method for: Data Quality Planning*, MP-17, EG&G Idaho, Idaho Falls, Idaho, June 1990, p. 8.